Developing a UMLS-based Ontology of Cardiology Procedures for Cognitive Support in Medical Decision Making

Jorge Biolchini, MD

Medical Informatics Training Program

Lister Hill National Center for Biomedical Communications

National Library of Medicine

- April 4, 2002 -

FINAL REPORT:

Title:

Developing a UMLS-based Ontology of Cardiology Procedures for Cognitive Support in Medical Decision Making

Author:

Jorge Biolchini, MD

Abstract:

Modularity of mental processes is a structural and functional characteristic of their organizational complexity.
Medical reasoning and decision making are conducted in specific, parallel and interconnected process levels.
Ontology is a formal conceptualization of a particular knowledge about the world, through the explicit representation of basic concepts, relations, and inference rules about themselves.
Domain ontologies can be used to provide knowledge support in underlying cognitive processes and inter-relations, and a methodology for connecting databases and facilitating professional communication, by supplying added-value information about structural and logical properties of the modeled conceptual network.
An ontology of the cardiology procedure domain was developed, based on information provided by the Unified Medical Language System.
Both intensional and extensional dimensions of UMLS concepts were taken into consideration.
Semantic contents of cardiology procedure concepts definitions, and conceptual relations in the Metathesaurus and the Semantic Network were used in the research, with additional information provided by MeSH Journal Descriptors.
Two methods were used for defining the subset of cardiology concepts, one lexical-based and another hierarchical-based.
The resulting set was revised to exclude concepts not strictly pertinent to the specific study domain.
The final definition set was submitted to natural language processing, analyzing phrases according to their basic elements and mapping the component concepts to UMLS 'Semantic Types'.
Conceptual categories were generated, and populated with the concepts, by using both a top-down and a bottom-up approach, and by defining a meronymic compositional structure for representing domain knowledge.
Relations between classes were defined, according to two approaches, a formal-rationalistic and an empirical one.
Tools used for developing the research consisted in a DBMS, the MetaMap, the Semantic Navigator, the Protégé platform for ontology development, and a graphic design tool.
The methodology used for developing the work will be presented, as well as the produced results.
These will be displayed in the specific developed taxonomy, and in Sowa's Conceptual Graphs to represent some of the associative relations.

INTRODUCTION AND BACKGROUND:
Cognition, Knowledge Representation, and Ontology

Cognition:

Modularity of mental processes is a structural and functional characteristic of their organizational complexity.
Both in terms of biological evolution as well as neuronal structures and physiological processes, human brain can be understood as a multi-organ system. Competing and cooperating kinds of interaction between neuronal regions and centers, working in a parallel distributed multi-level processing can help to explain some of the mechanisms and behavioral aspects of mental and cognitive processes.
In the last two to three decades the emerging field of cognitive sciences has developed and helped to bridge many interdisciplinary links to address the problem of human thinking, understanding, information processing, knowledge representation and correlated issues.
Since the beginning of researches, one of the major areas of cognitive studies has been the medical field.
In the first years of its scientific development, most of the studies have been conducted in the laboratory setting, following the dominant scientific paradigm of closed universe experimentalism.
"In recent years there has been a growing body of research examining cognitive issues in naturalistic medical settings." (1)
Due to the complexity of the issues involved in the human cognition phenomena and their dependence to environmental conditions, as well as to human interaction and human variability, new studies of the work place were developed in order to address, in a broader experimental perspective, subjects such as dynamic decision making, complex problem solving, human factors, and cognitive engineering.
"Current perspective on distributed thinking has shifted the onus of cognition from being the unique province of the individual to being distributed across social and technological contexts." (1)

In order to understand some processes about the way that human beings learn and perform their activities, many studies in this field have been based in the expertise-novice paradigm.
The rationale for this is to use the spectrum upon which a differentiation of individuals can be defined, according to their amount of knowledge and experience in a certain professional area, by using a discrete scale where distinct levels of expertise are represented, in order to provide information about their relative cognitive differences.
"Expertise-novice paradigm contrasts individuals in different varying levels of competency and training in order to characterize differences in cognitive processes (e.g., reasoning strategies, memory) and knowledge organization.
Typically, three levels of expertise are distinguished: novice, intermediate and experts." (1)
These studies, comparing professionals of different scientific and performing fields, have provided consistent results about the expertise-novice spectral differences.
As in other professional areas, medical reasoning and decision making processes are conducted through specific, parallel and interconnected levels. These processes are realized through series of shifts between these various levels of mental processing of information and knowledge. Sometimes these sequences include parallel simultaneous occurrence of different processing levels that in the end can either merge together in new sequences of other processes in a cooperative way. Or they can get into conflicts and consequently give rise to different and

diverging subsequent pathways, where sometimes one or more of them are interrupted or eliminated in order to resolve their competition.

Each one of these levels includes different sets of distinct mental operations with information, which can use a variety of representation modes for their execution.

Basically four major distinct and inter-related levels of cognition have been used as references for knowledge and competency analysis of professionals in different domains. They can be considered as cognitive-based meta-levels of information organization for professional purposes: factual, semantic, schematic and strategic. (2)

Knowledge Representation:

"Like Socrates, knowledge engineers and system analysts play the role of midwife in bringing knowledge forth and making it explicit. … To make the hidden knowledge accessible to the computer, knowledge-based systems and object-oriented systems are built around declarative languages whose form of expression is closer to human languages. Such systems help the programmers and knowledge engineers to reflect on "the treasures contained in the knowledge" (*Immanuel Kant*) and express it in a form that both the humans and the computers can understand." (3)

"Knowledge-based systems emphasize meaning. Instead of processing data as a string of bits, they represent the meaning of data in terms of the real world. They carry on conversations with people in ordinary language, they find important facts before they are requested, and they solve complex problems at expert levels of performance. … Two fields devoted to knowledge-based systems are cognitive science and artificial intelligence. Cognitive science is a merger of philosophy, linguistics, and psychology with a strong influence from computer science. Artificial intelligence (AI) is the engineering counterpart." (4)

"Knowledge representation is a multidisciplinary subject that applies theories and techniques from three other fields:

1. Logic provides the formal structure and rules of inference.
2. Ontology defines the kinds of things that exist in the application domain.
3. Computation supports the applications that distinguish knowledge representation from pure philosophy.

Without logic, a knowledge representation is vague, with no criteria for determining whether statements are redundant or contradictory. Without ontology, the terms and symbols are ill-defined, confused, and confusing. And without computable models, the logic and ontology cannot be implemented in computer programs.

Knowledge representation is the application of logic and ontology to the task of constructing computable models for some domain." (3)

Ontology and the medical domain:

Ontology is a formal conceptualization of a particular universe of knowledge about the world, through the explicit representation of its basic concepts, the relations between them, and some of the rules that provide inference about themselves and the things that they represent.

Domain ontologies can be used to provide knowledge support in some aspects of the underlying cognitive processes and their inter-relations. Another aspect of their use is that they can serve to provide a methodology for connecting different databases through the use of common knowledge structures. They can also provide an efficient means for facilitating professional communication about specific domain subjects and tasks. This is realized by supplying added-value information about structural and logical properties of the modeled conceptual network, and through the

explicit formal representation of the core domain concepts and their hierarchical and associative relationships, logically and ontologically constrained.
Previous researches in the development of ontologies for medical knowledge representation have highly relied in knowledge elicitation from domain experts. Most of these researches have also developed many parts of their work from the scratch. These two aspects have as a consequence a significant demand of human expert work.

UMLS-based Ontology of Cardiology Procedures

The objective of this research is to construct an ontology of different facets that are represented in the knowledge about medical procedures in the domain of cardiology, in order to build an information model for cognitive support in medical decision making processes.
The development approach proposed for the research is to use, for the purpose of its knowledge acquisition, the information and knowledge that can be found in the Unified Medical Language System (UMLS), developed in the National Library of Medicine.
Another aspect of the research goal is to emphasize the cognitive properties of medical decision and medical education processes as relevant guidelines for the knowledge organization and granularity levels to be achieved.
The reasons for choosing UMLS as a primary source include different features of it as a language system for medicine.
Integrating more than 60 vocabularies and classifications, some of them in multiple editions,(5) the UMLS constitutes an important knowledge *corpus* based on different knowledge sources.
Its inclusion of many different families of vocabularies accounts for a vast comprehensiveness in contents, which provide an important repository of concepts and their inter-relationships, in different levels of abstraction, allowing for their reuse in different clinical applications.(6)
The integration of these sources and vocabularies into one same meta-vocabulary makes it easy for using it as an inter-lingua for specialists in the medical field.
Its computability aspects, which include the data, their organization in files and tables, the many information links through different kinds of association, and the applications built upon it, such as browsers, Natural Language Processing tools, and others, represent an important set of resources that can be utilized for building new information and knowledge structures as well as to reproduce the existing ones by reorganizing them in different modes.

The ontology to be constructed in this research must be able to formally describe the most relevant contents and structures of information and knowledge related to the domain of cardiology procedures, in order to help in the design of applications based in the relevant cognitive components in this area. Fields for application include different aspects of medical decision making support and medical education.
In order to accomplish this goal, the ontology must be able to provide answers to the following competency questions: (7)
What are the information contents that are able to define a procedure in cardiology?
What are the agents and objects pertinent to a cardiology procedure?
What are the processes involved in the performance of a cardiology procedure?
What are the specific goals of a cardiology procedure?
What are the specific means used by a cardiology procedure?
How are the agents and objects related between themselves?
How are the processes related to the agents and objects?

The resulting ontology must be capable to reproduce in a formalized and computable way some of the relevant mental elements and relations that are present in the medical reasoning processes.

It must be able to formally represent and help to identify what elements constitute medical facts in the particular domain (factual level), what meanings are attached to these facts in terms of their different denotations and connotations (semantic level), and what structures of links between different meaningful elements reproduce some of the schemas that are learned and used for reasoning by professionals (schematic level). The only cognitive level that this ontology is not supposed to provide elements for is the strategic one. For this specific cognitive level, other elements and kinds of information and relationship are necessary, and can be better provided by other medical knowledge bases and sources.

In order to accomplish these knowledge representation goals, the ontology must be organized according to the basic principles of a conceptual meronymy, from which axes with specific taxonomies can be developed. The components of the medical act of a cardiology procedure must be represented in the highest specificity as possible, in a compositional mode, as parts of a whole, which is constituted by the procedure itself.(8) The degree of granularity of the categorial division, as well as of the distinct taxonomies associated with each axe, must be the highest as possible, according to the conceptual material that can be extracted from the UMLS, in order to make it feasible that the semantic categorization and structures generated in this process of knowledge representation can further be utilized for cognitive purposes in clinical applications and settings.

Although not constituting a primary goal of the research, developing this ontology can also be a means to help identify some lacks of information in the concepts of UMLS as well as the lack of some possible relationships between some of the concepts.

By working within the intensional conceptual dimension of UMLS, this approach can contribute to its enhancement as a unified language system for the medical field by providing a canonical formalized structure for refining the existent definitions of concepts.

By structurally and semantically comparing the elements contained in the intensional definitions, the identification of these lacks can become more evident, and display the pieces of information that could be added to each conceptual definition in particular, in order to make each one of them more comprehensive, informative, formally explicit, and consistent with the whole set of definitions in the specific domain.

As a canonical formalized and explicit structure of concept definition for procedures, it can also contribute to extend these information properties to other medical procedure domains.

The development of the methodology for using the UMLS as the basis for constructing this ontology is multifold in its purposes.

The methodological approach must be able to extract the maximum as possible of information contents from the concepts pertinent to the domain to be analyzed.

It must also make use of the information contained in the network of concepts, built in both the lower abstraction level, provided by the UMLS Metathesaurus, and in the higher abstraction level, as present in the UMLS Semantic Network.

The methodological approach must also be able to generate the maximum of coherent knowledge structures within the domains of cardiology in particular, as well as within medical reasoning as a whole.

These knowledge structures must be logically consistent and ontologically plausible, in order to be capable to reproduce the cognitive levels of reasoning processes involved in the medical activity, adding value then to the original information.

The methodology must also look for maximizing the possibilities of introduction of automatic phases in the ontology generation process, and consequently diminishing the human work component necessary to accomplish it.

As aforementioned, other methodological approaches have primarily relied in the human information extraction, by the process of elicitation of knowledge from experts in the domain.

Although this represents a possible approach for building ontologies, the author's methodological option privileges the reuse of the information and knowledge that is already present in the UMLS,

supported by one of the fundamental principles that constitute the utility of ontologies, extensively emphasized in the ontology community: the maximization of reuse of the existing information elements from databases and knowledge-bases.

METHODOLOGY

The methodology applied in the research started from two parallel processes for generating the initial data, and subsequently followed with a series of phases in a sequential mode of development, until reaching two different final forms of knowledge representation of the domain. The diagrammatic representation of the whole process developed in the methodology is displayed in Graph 1.

From domain words to domain concepts, and from seed concepts to domain concepts:

The definition contents of UMLS cardiology procedure concepts, as well as the conceptual relations present in the Metathesaurus and in the Semantic Network, were used as the basic material for the research.

The starting point of the methodology consisted in generating a subset of UMLS concepts containing definitions.

In order to meet these goals, the first step was to determine a subset of UMLS concepts, according to the following criteria:

The concepts should have definitions associated with them in the Metathesaurus, in order for them to be able to provide the necessary amount and specificity of information about the domain. This information would subsequently be extracted from the various contents of the retrieved concept definition texts. Treating these extracted definition contents as new information units would therefore allow the increase of both the amount and the diversity of conceptual elements and the consequent degree of the information granularity.

Concepts should also have either 'Diagnostic Procedures' or 'Therapeutic or Preventive Procedures' UMLS Semantic Types associated with them.

They should also be restricted to the specific domain of cardiology.

Concepts with 'Laboratory Procedures' Semantic Type were not considered for this study, because, in comparison to the aforementioned two selected Semantic Types, their concept definitions present important differences in terms of information contents.

Two alternative approaches were used to generate this subset of cardiology procedure concepts containing definitions.

One of the approaches was based in the lexical use of words and word strings with significant meaning for the research object of study, cardiology procedures. The rationale for using these words was based on the fact that cardiology procedure concept definitions should in principle contain textual elements that refer both to the sub-domain of cardiology and to the sub-domain of procedures. Both the classical differentiation of specialties in medicine, and the consequent number of medical specialties that are presently most extensive, are based on the anatomical criterion of classification. Therefore words that are pertinent to the anatomical entities that are objects of study in the field of cardiology should in principle be present in concept definitions of this sub-domain. On the other hand, procedures are medical acts. As a transversal axe of conceptual classification, it can be found to be present in many specialties in the health field. As other transversal axes, this is an independent variable in relation to the anatomical region of the body to which it might refer. Therefore concepts of procedures should also in principle contain words that are relative to actions performed in the medical acts that they represent.

Based on these assumptions, two groups of words were used as inclusion criteria to retrieve the definitions of cardiology procedures. Since some of these words present common lexical roots, in order to reduce the redundancy of some sub-groups and the number of necessary elements for retrieval purposes, some of these words were merged together into their corresponding lexical root string, which would consequently enable to retrieve all its lexical variations.

Words and lexical root strings related to various macroscopic anatomical parts of the cardiovascular system should then be combined with words and lexical root strings related to actions which occur in cardiology procedures, related either to diagnostic and to therapeutic or preventive types.

For defining a more efficient group of words related to the anatomical entities of the cardiology domain, additional information was provided by the Medical Subject Heading's Journal Descriptors, in terms of their relative frequency as found in the medical literature.

Terms with higher frequency were then considered to have a potentially stronger retrieval power for cardiology concept definitions.

The Medical Subject Heading's Journal Descriptors were also a relevant source, not only for "suggesting" new words for retrieval purposes, but for two other goals. Checking one by one if each element of the set of anatomy-related words could really be helpful for retrieval purposes, and in what measure, including each associated "noise". It was also an important resource for optimizing the character extension for each string. Since one of the specific objectives consisted in optimizing the automation of the process, and since some of the chosen words have lexical variants, some strings were defined in order to contain just the word root part. For others words this was not done. The Medical Subject Heading's Journal Descriptors helped to define which should be reduced to their root string and which ones should not.

For instance, for the variants 'Atrium', 'Atrial' and 'Atrio', separate strings were used for each one of these variants. Because if 'Atri' was used, in this case, a set of irrelevant words would also be retrieved (besides their other variants too), such as:

'Natriuretic', 'Matrix', 'Pediatric', 'Cicatrix', 'Psychiatric', '8, 11, 14-Eicosatrienoic acid', 'Pregnatrienes', 'Diatrizoate', 'Sumatriptan', 'Geriatric', 'Veratridine', and 'Veratrine'.

The tables of the 2001 version of UMLS were imported into a Data Base Management System, the major source of information being constituted by the MRDEF table.

From the group of cardiology anatomy related words and lexical root strings a query was executed and a first subset of UMLS concepts was generated, containing 2929 concepts.

From this subset, a group of words and lexical strings related to specific procedure actions in the domain of cardiology was generated.

Since, from one hand, the concept definitions are not formally defined, being intensely heterogeneous both in relation to their semantic contents and to their linguistic structure, and since, from the other hand, there are overlaps between the information contents of different medical sub-domains, concepts not pertinent to cardiology procedures would certainly be found among the retrieved ones. Therefore restriction criteria should be applied to the final query in order to restrict the sub-set of concepts to the minimum number as possible, as well as to the more specific ones. As a consequence of this restriction, the human review process of the final retrieved set could also be minimized.

The set of strings used to exclude irrelevant concepts for the field of cardiology procedures was generated through human review of the initial query in the MRDEF table, using the anatomy set of strings.

For restricting qualitatively and quantitatively the sub-set of concepts, words referent to other anatomical entities, as well as to medical instruments that do not concern to the cardiology procedure domain, were used as exclusion lexical factors.

To the MRDEF file a final query was applied, containing the three sets of strings mentioned above, in order to retrieve the specific concepts of cardiology procedures.

An example of such a combination of the three sets of words and strings would be using as one of the anatomy strings ‘heart’, as one of the action strings ‘inject’, and as one of the exclusion strings ‘trachea’. The first two strings would be able to retrieve, as they did, the following concept definition, where the exclusion string is not present. The bold letters are included by the author in order to highlight the referred strings:
“Radiography of the **heart** and great vessels after **injection** of a contrast medium.”
A little group of these retrieved concepts contain more than one definition associated with each one of them, that are brought by the original vocabularies from which the concepts were incorporated into the Metathesaurus. Although some of these plural definitions for singular concepts are very similar to each other, others are not, and may contain different elements of information. For the purposes of the research, the original information unit of interest was defined as being the concept itself. The plural association of definitions was therefore considered as if it was equivalent to an enriched definition for each of these same concepts in terms of contents. The redundancy of equivalent elements in each plural association of definitions for one same concept was further eliminated in their content analysis.

The second approach used to generate a subset of UMLS concepts containing definitions was based in the hierarchical relationships that exist between concepts in the Metathesaurus. The rationale for using these relationships was based on the fact that cardiology procedure concepts should in principle be linked through broader concepts in the taxonomical structure of the Metathesaurus. Because of the multi-axial semantic links of the concepts in the hierarchy, some sub-groups of them could in principle be dispersed in different conceptual regions of the whole network of the Metathesaurus nodes. Therefore using concepts in the domain of cardiology that could function as seeds for finding their hypernyms and siblings could in principle be a capable means for retrieving a bigger number of related concepts, through the process of extending the resulting set to their conceptual neighborhood nodes. These seed concepts also needed to be able to represent both diagnostic and therapeutic or preventive aspects of cardiology procedures, as well as both clinical and surgical modalities of medical intervention.
The major direction to be followed, in this migration path through correlated concepts, should then be the initial search for good seeds, in order to use their whole potential to identify a network of other nodes, in two sequential stages. The first step would start from the bottom-up in the hierarchy, in order to look for their corresponding hypernyms. The second stage would then consist, from the top-down, in gathering together all the corresponding hypernym hyponyms. Possible overlaps of equivalent concepts coming from different hypernyms should also be eliminated through their corresponding unique identifiers in the system. Good seed, in the context of the present research, means unequivocal cardiology procedure concept, having as a consequence of this definition its more precise localization as a node in the network of linked concepts. Empirical pilot validation of the seed value of an initial set of good candidates for the purpose of retrieving satisfactory hypernyms, in terms of their extension degree of coverage of linked hyponyms, was realized by using the Semantic Navigator.
The concepts that were finally defined as actual comprehensive nodes for retrieving cardiology procedure concepts were the following, including their unique identifiers:

- C0038897 | Cardiovascular Surgical Procedures
- C0189572 | CARDIOVASCULAR SYSTEM: GENERAL AND MISCELLANEOUS OPERATIVE PROCEDURES
- C0011904 | Diagnostic Techniques, Cardiovascular
- C0199533 | MEDICAL PROCEDURES ON THE CARDIOVASCULAR SYSTEM

Four subsets of concepts were then retrieved and joined together into one final set.

These two approaches, lexical-based and hierarchical-based were conducted separately and in parallel, and were later on compared in relation to their corresponding sets of retrieved concepts.
A human review process was conducted over the retrieved sets in order to further select the specific sub-set of strict cardiology procedure concepts.

From domain concepts to domain semantics:

The final set of cardiology procedure concepts containing definitions constituted the starting point for the next phase of the methodology.
The objective in this step was to identify the various semantic elements contained in the definition texts of the selected concepts.

The whole set of textual definitions of the concepts was analyzed through Natural Language Processing by using the MetaMap Technology of the Semantic Knowledge Representation project of the National Library of Medicine.
The parameter options that were used for language processing included:
- Mappings (-m)
- Best Mappings Only (-b)
- No Acronym/Abbreviation Variants (-a)
- Semantic Types (-s)
- SemRep Output (-S)

The Output Text was further organized and displayed in different formats.
Since some of the mapped terms, contained in the textual phrases of the definitions, have more than one Semantic Type associated with each one of them, the first output format consisted in separating the different concepts, composed by pairs containing a term associated with a specific Semantic Type. Treating them as independent units could then provide full use of the different connotations of an isolated term, in the semantic context of the UMLS Semantic Network, and allow us to make a controlled use of the terminological polysemy of the isolated terms, after being taken out of their original contexts of the definition sentences.
This output format was then submitted to human review in order to identify, according to the specific contexts of the textual concept definitions from which they were derived, those concepts which meaning was coherent to the cardiology procedure domain and to differentiate them from those which meaning was not coherent to this context.
The MetaMap equivalent concepts, extracted from different definitions, were further fused together into one same concept for all its repeated extracted elements, obtained from the mapping process.

The second output format consisted in grouping together the concepts, according to their corresponding Semantic Types, in order to allow an analysis of the full content and semantic variability of the mapped terms, in reference to each one of the different associated Semantic Types. By doing so, the number of concepts per each Semantic Type could be treated as a directly proportional indicator of the semantic relevance of each Semantic Type in the context of the cardiology procedure domain.

From domain semantics to domain ontology:

After realizing this process of ontology capture of the relevant elements and semantically organizing them, the next step should then be the definition of the basic categories for representing the different information components of a procedure.

The conceptualization of the fundamental units for the adequate knowledge representation of a procedure in the domain of cardiology, as well as the specification of the conceptual elements responsible for containing the different and related entities in this representation, was realized by using both a bottom-up and a top-down approaches.

The bottom-up approach was done by analyzing and comparing the semantic components, which were generated by the mapping process, both inside each of the Semantic Type groups and between those groups that contained concepts that presented some kind of semantic relationship between them.

This comparative analysis was followed by grouping the Semantic Types together in categories, according to these similarities of meaning of concepts and groups of concepts.

This grouping process had as a reference the competency questions that the ontology was intended to cover, in order to be able to better describe the universe model that it should be designed to represent.

The top-down approach was referenced to a basic schema of the agents and processes that interact in the real world, in the setting of the execution of a general medical act.

This schema can be found in Graph 4.

A comparison of the Semantic Types and their corresponding concepts analyzed by MetaMap was then realized with the types of elements that are present in this general medical act schema.

This comparative analysis was then realized in relation to the former information model designed by the author prior to this present research, which has been based in a small sample of UMLS cardiology procedure concept definition contents.

This categorization process had, as a conceptual frame of reference, the objectives of the research and the need to develop a knowledge representation of the semantic components of a complete formal meronymic definition, designed to include the information contents that allow to describe any cardiology diagnostic and therapeutic or preventive procedure.

The final categorization was compared to the Semantic Groups of the UMLS Semantic Types for improving the categorial consistency.

Finally the categories were also compared to the MAOUSSC axes for representing medical procedures.(9)

These comparisons were able to provide a more precise definition of the categories needed to represent the medical procedures in the domain of cardiology according to the granularity level expected for this research and its cognitive objectives.

The methodological use of the association of both top-down and bottom-up approaches constituted then what some authors call the middle-out approach of ontology development.

This process gave origin to a group of basic categories, in order to match the different mapped concepts and define their corresponding place in the basic structure.

These categories were then separated in two super-categories, according to their compositional relevance to represent a procedure. The first super-category was composed by a group of so-called nuclear categories, and the second one was composed by another group of so-named auxiliary categories.

The classification of the basic categories in these two super-classes was two-fold motivated.

On the one hand, it could help to identify, in a higher level of abstraction, distinct areas of relevance, in a general sense, for the representation of the information elements of the categories. By doing so, the nuclear classes can represent more essential characteristics and intrinsic properties of what constitutes a procedure, as well as the basic information elements that are commonly thought about medical procedures among professionals and students. At the same time, the auxiliary classes might be able to represent more complementary features and extrinsic properties for the thinking process about procedures. Therefore this higher-level grouping of categories would be serving to a cognitive function.

On the other hand, according to the canons of classification theory and methodology, it could help to produce a more elegant and well-distributed arrangement of the different types of

elements in the whole tree structure. Subjacent to this theoretical and methodological canon of class distribution, a cognitive motivation can also be recognized, in the sense that it helps to improve the visualization of the whole classification structure, as well as to enhance the quality of the consequent apprehension process of the meanings for the particular designed arrangement of the classes.

This is considered to be one of the two general and basic principles for the formation of categories. It "has to do with the function of category systems and asserts that the task of category systems is to provide maximum information with the least cognitive effort." (10) Named as cognitive economy principle, it "contains the almost common-sense notion that, as an organism, what one wishes to gain from one's categories is a great deal of information about the environment while conserving finite resources as much as possible." (10)

Once having the categories defined, the following step consisted in defining this group of categories as a fundamental compositional structure of the conceptual elements contained in any of the cardiology procedure concept definitions. This compositional structure could consequently be used to formally explicit a canonical descriptive meronymic definition that could be valid for the whole set of UMLS cardiology procedure concepts. In natural language this definition could then be formulated as the following:

(*the names of the categories are written in bold letters*)

"A cardiology procedure concept definition is a conceptual entity, which is composed by an **action**, which: has location in an **anatomical entity**, and affects a **pathology**, and affects the **physiology**, and has a **purpose**, and uses an **instrument**, and uses some **material or energy**, and can be preceded by a **conditional action**, and has a **method**, and uses or affects a **phenomenon**, and uses a **measure**, and has a **result or product**, and occurs in a certain circumstance in reference to **time**, and has some **space** references, and affects a **receiver**."

Because the very meaning of 'procedure', derived from the latin junction of *pro-* (forward) and *cedere* (to go) (11), is of an active process ('to go forward'), essentially implying an action, the innermost nuclear component of a procedure is the category that corresponds to the action itself that is used to accomplish the procedure.

Therefore, except for the conditional action, all the relations that the class action has with the other categories are transitive to the procedure itself as a conceptual entity and category. Since both the action and the conditional action are components of the procedure, with a vertical part-whole relation kind between each one of them and the procedure, and since, at the same time, the conditional action is in itself another kind of action, the possible relations of these last two components of the procedure between themselves are, on the one hand, of a temporally related associative kind ('precedes') and, on the other hand, of a horizontal hierarchical kind (as siblings).

Having the information model defined, the next step of the research was to develop, for each category of the syntagmatic plane of the compositional definition structure, the conceptual network of the pertinent elements, in their paradigmatic plane.

This was a tailor-made process of definition and sub-categorization, in order to characterize the subtypes and the hierarchy corresponding to each definitional facet of the procedure conceptual structure.

By using a bottom-up approach, this process was realized according to the available set of concepts generated by the previous phases of the research, and associated with each of the basic categories.

The frame of reference for this categorization was the main objective of the research, according to which the categorical structures should, in principle, be able to represent the domain concepts and relations for cognitive purposes. Therefore it should be able to attend the following requisites:

The differentiation of distinct facts and qualifiers of facts, on the one hand, and the union of similar kinds of facts and qualifiers of facts, on the other hand, from the medical knowledge perspective, should be represented in corresponding classes and levels of specification-

generalization. For instance, the concept ‘Severity’ was located under the category ‘Pathology’ as a ‘concept of pathology modifier’, while the concept ‘Arrhythmia’ was placed in a sub-category named ‘Specific disease’ under the same category ‘Pathology’.

The classes should be able to semantically differentiate and, at the same time, to group the most significant features of the concepts for each particular class, with the perspective of having them represent elements and parts of an active process (*pro-* + *cedere*), as executed by professionals, as their top level whole, and not by their meaning in themselves as isolated units from this whole.

Therefore, following the example above, because of both characteristics of the concept ‘Severity’, that is, as a modifier, and also as a concept in the higher categorical level, ‘Severity’ has a transitive property in relation to ‘Arrhythmia’, and both concepts can be joined in a composite expression such as ‘Severity of Arrhythmia’, or else be further derived into ‘Severe Arrhythmia’.

The third requisite was that the sub-division of different categories should achieve some parallelisms in structure, both in relation to the basic characteristics of the specific universe of the world, which they represent, and in relation to other categorical sub-structures. This dual consistency, external and internal, would approximate them to some of the schematic organizations that exist in the medical field, as well as facilitate the most direct establishment of relations between some of these sub-classes. This approach would also be able to determine some restrictions in relationships between sub-classes, further enhancing the formal representation of the domain knowledge.

For instance, the concept ‘Balloon Dilatation’ was located in the sub-class of ‘Action’ named ‘Dilatation’, while the concept ‘Balloon dilatation catheter’ was located in the sub-class of ‘Instrument’ called ‘Mechanical instrument’.

Another example is that the concept ‘Electric potential’ was located in the sub-class of ‘Phenomenon’ named ‘Electrical phenomenon’ while the concept ‘Electrocardiographic recorders’ was located in the sub-class of ‘Instrument’ called ‘Electrical instrument’.

The next step of the research consisted in populating the distinct categories with the different concepts.

The variance in connotation that the concepts represent, by having in some cases the same term associated with different Semantic Types, is able to provide a richer semantic representation of definition contents in the organization of the separate classes. Since the terms were now isolated from their original textual contexts, they have suffered a certain loss of information in this detachment from their original network of relations within the textual neighborhood of information elements. Inserting them in the new context, produced by the categorical structure of representation of definitions, semantically means to add to them a new and restructured network of relations with other information units in a more abstract level. Since all reminiscent pairs of ‘term-Semantic Type’ were considered to be valid information units in relation to their original textual context, inserting them in the new compositional information structure could mean that, for some of the different pairs associated with the same original term, each one of them could then possibly establish stronger relationships of meaning with different categories, while still preserving a link to the original context as a valid signification of the associated original term.

One example can be the term ‘*Airway*’, which is associated with two different Semantic Types, ‘Body Space or Junction’ (bsoj) and ‘Medical Device’ (medd). In the original textual contexts, both were valid concepts. In the categorical structure and populating process, the first Semantic Type matches with the ‘Anatomical entity’ category, while the second one is linked to the ‘Instrument’ category.

A different case occurs with the term ‘*Animal*’, which has two different Semantic Types associated with it, ‘Animal’ (anim) and ‘Group’ (grup), both valid in relation to their original contexts. Here both Semantic Types are related to the same category of ‘Receiver’.

A more complex case is the term ‘*Blood pressure*’, which is associated with three different Semantic Types: ‘Diagnostic Procedure’ (diap) through the concept ‘*Blood pressure determination*’, ‘Laboratory or Test Result’ (lbtr) through the concept ‘*Arterial pressure*’, and

'Organism Function' (orgf) through the concept '*Blood pressure*', all valid in their original contexts. While the first Semantic Type correlate to the category 'Action', the second one is related to the category 'Result or product', and the last one corresponds to the category 'Physiology'.
These different locations for concepts in the categorical organization not only preserve the original meaning of the concepts, but also provide a means to establish direct relationships between the classes through their corresponding conceptual elements.
Therefore, in the first case, '*Airway*' can be defined as an 'Instrument', which is related to the 'Anatomical entity' '*Airway*', relationship which is constrained through an 'Action' of 'Substitution' (a sub-class of Action).
In the third case, '*Blood pressure*' as '*Blood pressure determination*' is an 'Action', which acts upon '*Blood pressure*' in 'Physiology', and produces '*Arterial pressure*' as a 'Result or product'.

The populating process was realized in two independent ways, so that in the end they could be compared.
The first method consisted in manually populating the classes with the concepts. The second method corresponded to generate an automatic populating process of the categories.
The motivation for doing so was two-fold. On the one hand, automating this phase of the process, as in other steps of the methodology, could bring as a consequence the easier generalizability of the methodology, and then make it possible to extend it to other medical domains. On the other hand, making a comparison between both methods, by having assumed the human-based method as the gold standard for the automatic one, could provide a means for evaluating the results of the automatic populating process.
The only Semantic Types that were not included in the automatic populating process of the categories were 'Functional Concept' (ftcn), and 'Qualitative Concept' (qlco). Because of their high concept diversity, both were able to generate a group of sub-types for each one of them, matching then in the ontology a corresponding number of different categories and sub-categories. The first one gave origin to 14 different sub-types, while the second one to 12 different sub-types.

The next phase of development was to construct a knowledge representation for the categorical structure populated with concepts.
A knowledge edition environment was used for this task, the Protégé platform for building ontologies, developed in the Stanford University.
The conceptual structures were built in this frame-based system, with different slots being defined, in relation to each particular class, for each set of corresponding concepts. Other slots were further defined for incorporating the set of dyadic associative relationships established between each pair of the ontology classes.

<u>From categories to relations</u>

The next phase of the research methodology was to establish the different kinds of possible associative relations between the categories.
Two approaches were used for accomplishing this task. The first approach was a formal-rationalistic one, while the second was a formal-empirical one.
The set of the 53 associative semantic relations of the UMLS Semantic Network served as a conceptual reference for the kinds of potential joints that could be made between the classes. This reference set was utilized for both of the approaches.
The difference between the approaches consisted in both the starting material and the direction of the process.

The formal-rationalistic approach used the group of basic categories and their corresponding sub-categories in order to develop the ontology capture of the relationships

between them. It started by the identification of the possibilities of associative relations for each pair of classes, and used a top-down direction for this process.
Sowa's model of conceptual structures provided an important reference for this approach, both in terms of the set of relations to be thought about, as well as in terms of the syntactic formal expressions to be used in order to represent the produced relations. (4)
The constructed relationships were defined, as well as some of the most relevant inverse relationships.
For each relation, primarily defined through a formal and abstract joint of classes, the substitution of the names of the two related arguments by some of the concepts associated with them was realized, in a second stage, in order to check for its plausibility in the concrete sphere of the referents of the concepts.

The second approach was a formal-empirical one. The working material was composed by the original textual definitions of the cardiology procedure concepts.
The set of semantic relations of UMLS Semantic Network served as a syntax model in order to formalize the natural language expressions for some of the relations that could be identified in the textual definitions between each pair of retrieved concepts. The main goal here was to be able to translate, from an expert perspective, the higher number of plausible relationships that could be found in the textual concept definitions, by using the Semantic Network semantic relations as their formal equivalent expressions.
In a second stage, for each identified semantic relationship built by this process, the same syntax was applied to the corresponding Semantic Types associated with each pair of concepts, having them as new arguments for the same relation, in order to check for the ontological consistency of the same relationship in a higher level of abstraction.
The restrictions for the relations as defined in the Semantic Network were not used as constraints for the results of this phase of the research. The constraints applied for this bottom-up approach were context-related to the domain, and were therefore constituted by the results of the ontology relation definitions as derived from the top-down approach.

Formal and graphical representation of the relations

The hierarchical relations between categories and some of the associative relations of the ontology were developed in the Protégé platform.
The last phase of the research consisted in formalizing the syntactic definition of the developed relations and constructing their corresponding graphical representation.
Sowa's conceptual graphs were used for both tasks.
In the first stage, the algebraic form of representation for the relations was done for each defined relationship, by using the elements of the conceptual graph knowledge representation model, such as its notation, syntax and types of formal relations.
According to the kind of relationship, different conceptual structures were then constructed, based on the involved elements, on the valence of the relation itself, and on the argument direction for each specific valence of the relation.
This symbolic coding served as an explicit and formalized representation of the conceptualization for each specific link between concepts in their abstract level.
For each of these relationship, symbolic linear structures and an equivalent diagrammatic representation was further developed.

RESULTS:

From domain words to domain concepts, and from seed concepts to domain concepts:

"The 2001 edition of the Metathesaurus includes about 800,000 concepts and 1.9 million concept names in different source vocabularies." (5)
The total number of Metathesaurus concepts with definitions associated with them is 34,095. Therefore, approximately 4.26 % of the Metathesaurus concepts have definitions.

The obtained results to generate the subsets of cardiology procedure concepts are presented in Table 1 and in the corresponding Graph 2.
The total number of retrieved concepts with associated definitions included 468 concepts.
The total number of retrieved concepts with associated definitions included 253 concepts.
Therefore, 1.37 % of the total number of Metathesaurus concepts with definitions was retrieved by one approach, and 0.74 % by the other one. This means that the retrieval process generated a set with approximately 1 % of the total number of Metathesaurus concepts with definitions.
The human review process selected 124 final concepts of cardiology procedure from the lexical-based set, among which corresponded 76 of the final concepts selected from the hierarchical-based set.
The results that were obtained are presented in Graph 3.
The total number of common concepts, retrieved by using both approaches, was equivalent to 99 concepts. From this group of common concepts, 76 concepts were specific of the cardiology procedure domain and 23 were not cardiology procedure concepts but pertinent to related medical domains such as angiology and hematology.
The final set of the 124 selected concepts represented then 0.36 % of the total number of Metathesaurus concepts with definitions. This number corresponds approximately to 0.000155 % of the total number of UMLS Metathesaurus concepts.

From domain concepts to domain semantics:

From these 124 selected concepts, the MetaMap Output Text gave origin to a total of 2152 extracted concepts.
Among the textual definitions processed by MetaMap, typographical errors were detected some of the definitions, as they are found in the Metathesaurus, such as some of the definitions starting the first word of the first sentence with lower case, which does not happened to affect the natural language processing executed. The other minor error that was found was one word cut in the middle of it ("adjus ting").
From the 2152 MetaMap extracted concepts, 987 concepts were considered to be correct in reference to their textual contexts in the definitions, based in the human review. This means that the 124 original concepts were able to give origin to a derived number of concepts which corresponds to 796 % of their original number as a set, that is, quantitatively increasing 8 times its information power. Another way to apprehend this information generative process is considering that each concept definition gave origin to a mean number of 17 concepts, which means that each concept definition contributed with 8 new concepts to the whole of the 987 mapped concepts.

The positive and negative results of all the MetaMap extracted concepts for the concepts in the definitions, in relation to their contextual coherence, are displayed in Table 2.
The results of the subsequent fusion of extracted concepts into unique concepts are displayed in Table 3.

The obtained results of the groupings of the second MetaMap output format, per each Semantic Type, and sorted by their descending number of terms, are displayed in Table 4.
From the original two Semantic Types 'Diagnostic Procedure' and 'Therapeutic or Preventive Procedure', 68 new Semantic Types of the total number of 134 were present in the associations with the selected concepts.

From domain semantics to domain ontology:

The comparison of the categorization process with the former information model designed by the author prior to this present research is found in Table 5.
The comparison of the final categorization with the Semantic Groups of the UMLS Semantic Types can be found in Table 6.
The comparison to the MAOUSSC axes for representing medical procedures is displayed in Table 7.

This middle-out approach for the ontology development gave origin to 15 basic categories, under which 97.1 percent of all the mapped concepts could find their place. The obtained results for each category are displayed in Table 8.
These 15 categories were then separated in two super-categories, according to their compositional relevance to represent a procedure. Quantitatively, the first one, composed by 7 so-called nuclear categories, comprised 68.2 percent of the concepts. The second group, composed by 8 auxiliary categories, included 31.8 percent of the concepts.

As afore mentioned, for the automatic populating process of the categories, both the Semantic Types 'Functional Concept' (ftcn) and 'Qualitative Concept' (qlco) were not included, due to their diversity of relations with the ontology categories, respectively 14 for the first one and 12 for the second one.
The compared results of both the manual and the automatic category populating methods can be found in Table 9.

Because of the hierarchical relationship between a category and its sub-categories, all the concepts associated with the former have transitive properties in relation to the hierarchically related sub-classes.
The only categorization that was not done by the author was the correspondent to the 'Anatomical entity' paradigmatic axe. Reusing the UMLS-based Digital Anatomist ontology has been the option for this specific procedure definition facet.
The categorization of the 'Action' class started from the basic dialectic division between diagnosis and therapeutics, also existent in the UMLS Semantic Network as subtypes of procedures. For the diagnostic axe, the next division followed the basic reasoning and classical principle of the different senses of perception of the human being that are used in the medical act of diagnosis. In this level, the sub-division of the visual sense got a sub-category, based in the technological dimension that is dominant in our days for exploiting the visual perception of the human body. For the therapeutic or preventive axe, the next division followed the dualistic mode of division that historically and cognitively underlies human reasoning about interventions. Two pairs of dual division were used in parallel, corresponding to the inner-outer space reasoning and to the high-low intensity one. A fifth sub-category, which cannot be reduced to these two pairs, related to the notion of life organization and consequent health reorganization, completed this level. Further sub-divisions for each of these five classes of concepts were based in the particular differentiating features of the concepts involved.
The categorization of the 'Pathology' class consisted in the basic differentiation of the medical semiologic reasoning and practice for 'sign or symptom' and 'disease'. The existing set of concepts retrieved from the concept definitions permitted a further differentiation for both sub-

categories, according to the specificity and non-specificity of the retrieved concept for each of the two referred classes.
The categorization of the 'Instrument' class was based in the physical properties of the devices and tools that are used in medical procedures. Based on this categorization principle, some of these sub-classes could establish a direct parallelism in meaning with sub-classes of other categories, fulfilling some of the requirements for this organization and representation process. More specifically, a cognitive economy of the basic principles needed for categorization, an ontological consistency with the represented world, an inner logical coherence between different axes of knowledge representation, and a cognitive economy for generating relationships between different axes, were consequences of this approach.
The categorization of the 'Material or Energy' class was based in the physical nature of the referents of the concepts as well as in the natural-artificial origin of them.
The categorization of the 'Physiology' class was based in the physical nature of the physiological processes.
The categorization of the 'Purpose' class consisted in differentiating the specificity and non-specificity of the retrieved concept. Most concepts correspond to verbs and adjectives, because of their original textual syntax role in the definitions.
The categorization of the 'Method' class also consisted in differentiating the specificity and non-specificity of the retrieved concept.
The categorization of the 'Conditional action' class received no further categorization, due to its small number of concepts.
The categorization of the 'Phenomenon' class was based in the physical nature of the phenomenon involved, with a parallelism with the other mentioned classes that also were sub-classified based on this criterion.
The categorization of the 'Measure' class relied in its component primitive notions of 'number' and 'dimension'.
The categorization of the 'Result or product' class was based in the top-level ontological differentiation of 'result or product' as either an entity or a process.
The categorization of the 'Time or circumstance' class was based in some the primitive notions of temporal representation, such as 'duration', 'frequency', 'phase', and 'origin'.
Coherently, the categorization of the 'Space' class was based in some the primitive notions of spatial representation, such as 'spatial location', 'spatial configuration', 'spatial limit', 'spatial change', and 'spatial relation'.
Finally, the 'Receiver' class was not sub-categorized, due to a small number of concepts associated with it.

From categories to relations

The total number of relations, constructed through the bottom-up approach, is 1,405.
An initial comparison of the results of both methods for generating relations gave origin to 96 basic relations which were consequently included in the ontology.
The analysis of the results developed through the bottom-up approach has not been completed yet, and will undergo further studies.

Formal and graphical representation of the relations

The overall structure of the developed cardiology procedure ontology, as developed in Protégé, is displayed in Figure 1.

A small relevant sample of the diagrammatic representation of the relationship symbolic linear structures, according to Sowa's knowledge representation model, can be found in the Graphs 5, 6, 7 and 8.

Graph 9 finally displays, in a summarized form, the conceptual extraction and enrichment processes realized during the research.

CONCLUSIONS AND DISCUSSION:

The objective proposed for this research has been achieved.
Although there seems to be an important difference in the number of concepts retrieved by both approaches, in absolute terms, human review done in both sets shows that the numbers tend to approximate in relative terms.
The retrieval process was not exhaustively done with different sets of lexical strings and with seed concepts, because the objective of this research was not directed to it. Nevertheless, since the number of concepts that can be retrieved by each approach can vary according to the number of elements used for them, this strongly suggests that by narrowing or broadening this independent variable the final numbers of retrieved concepts by both approaches can produce more proximate results.

The high percentage of the total positive results in the automatic method (98.18%) already shows a high rate of accuracy of the automation of this process. Looking to the false negatives (0.71%) gives us a clearer picture of the potentialities of the method.
The high percentage of the false positive results (17.43%) clearly demonstrates the strong contextual dependence of concept meanings.
If the automatic populating process is then filtered out of the incorrect concepts in relation to the specific contexts, by eliminating them before doing it, the method dramatically improves, and the total of positive results rises to 99.13% of the total number of concepts, leaving out 0.89% of them.
A further improvement of the method can be realized through the restriction of Semantic Types associated with the concepts.
Table 10 shows this through the number of Semantic Types that were found for each kind of result.
The H(+)-A(+) result has 64 Semantic Types associated with the concepts.
The H(+)-A(-) result has 44 Semantic Types associated with the concepts.
The H(-)-A(+) result has 4 Semantic Types associated with the concepts.
The H(-)-A(-) result has 4 Semantic Types associated with the concepts.
When each two pairs of the four different kinds of results are compared, in relation to their equivalences and differences of Semantic Types associated with the concepts, a confirmation of the high contextual dependence of concept meanings emerges again in another way.
These results are shown in table 11.
In the upper right region of the table are the numbers of equivalent Semantic Types between each two kinds of results. For this region the 3 following results were found:
a) The pairs:
H(+)-A(+)/H(+)-A(-), H(+)-A(+)/H(-)-A(-), H(+)-A(-)/H(-)-A(+), and H(-)-A(+)/H(-)-A(-)
have no equivalent Semantic Types in their concepts.
b) The pair H(+)-A(-)/H(-)-A(-) has only one equivalence.
c) While the pair H(+)-A(+)/H(-)-A(+) has 43 equivalent Semantic Types.

In the lower left region of the table are the numbers of different Semantic Types between each two kinds of results. For this region the 3 following results were found:
The pairs:
d) H(+)-A(-)/ H(+)-A(+), H(-)-A(+)/H(+)-A(-), H(-)-A(-)/H(+)-A(+) and H(-)-A(-)/H(-)-A(+) have 4 different Semantic Types between each kind of result.
e) The pair H(-)-A(-)/H(+)-A(-) has 3 different Semantic Types between each kind of result.
f) While the pair H(-)-A(+)/H(+)-A(+) has one different Semantic Type between each kind of result.

Through this comparison, it also becomes clear that a restriction of Semantic Types, executed either prior to the mapping process, or in a second stage, after it and prior to the automatic populating process, or even a combination of both restriction phases, could be an efficient way to eliminate the remaining negative results, reaching 100% of accuracy in the automatic method and reducing human work in the process.
In the particular example of this domain, if the previous washing out of the negative results is realized, prior to the automatic populating process, the two right columns and the two lower rows are eliminated from the table. The remaining pair has no equivalence of Semantic Types, having at the same time 4 different ones.
The Semantic Types which should be eliminated in this specific context of cardiology procedures before populating the categories are the following:
- Biomedical Occupation or Discipline (bmod)
- Human-caused Phenomenon or Process (hcpp)
- Regulation or Law (rnlw)
- Social Behavior (socb)

Future work will consist in improving some specificities of the methodology, as well as the development of applications based in this knowledge representation for the domain.
Further work can be done in extending this methodological approach to other medical domains, and comparing results obtained in other medical specialties.
Regarding the methodology itself, the objective would consist in increasing the efficiency of both retrieval methods, by making other tests, using as variables the scope of elements for retrieval, their relative frequencies and their relative extensions of coverage.
One possibility for increasing the automation process of the sub-categorization phase would consist in incorporating inside the categories the own existing Metathesaurus hierarchical relationships between the concepts. This would help to structure the different sub-categories for each particular category. Inheritance of the existent hierarchical relations of concepts inside the Metathesaurus would then function as a way to generate hierarchy inside the information model categories.
A range of applications can be devised by using the cardiology procedure ontology.
The afore mentioned areas related to medical decision making and medical education would benefit from the logical properties of the relations between categories populated with concepts, such as establishing chains of links in order to reproduce some properties of the cognitive processes involved in medical reasoning.
Other areas such as Natural Language Processing could also get some benefit by using the conceptual structures and the semantic features and constraints as parameters for improving identification and association of some elements.

Graph 1:

Diagram of the methodology phases

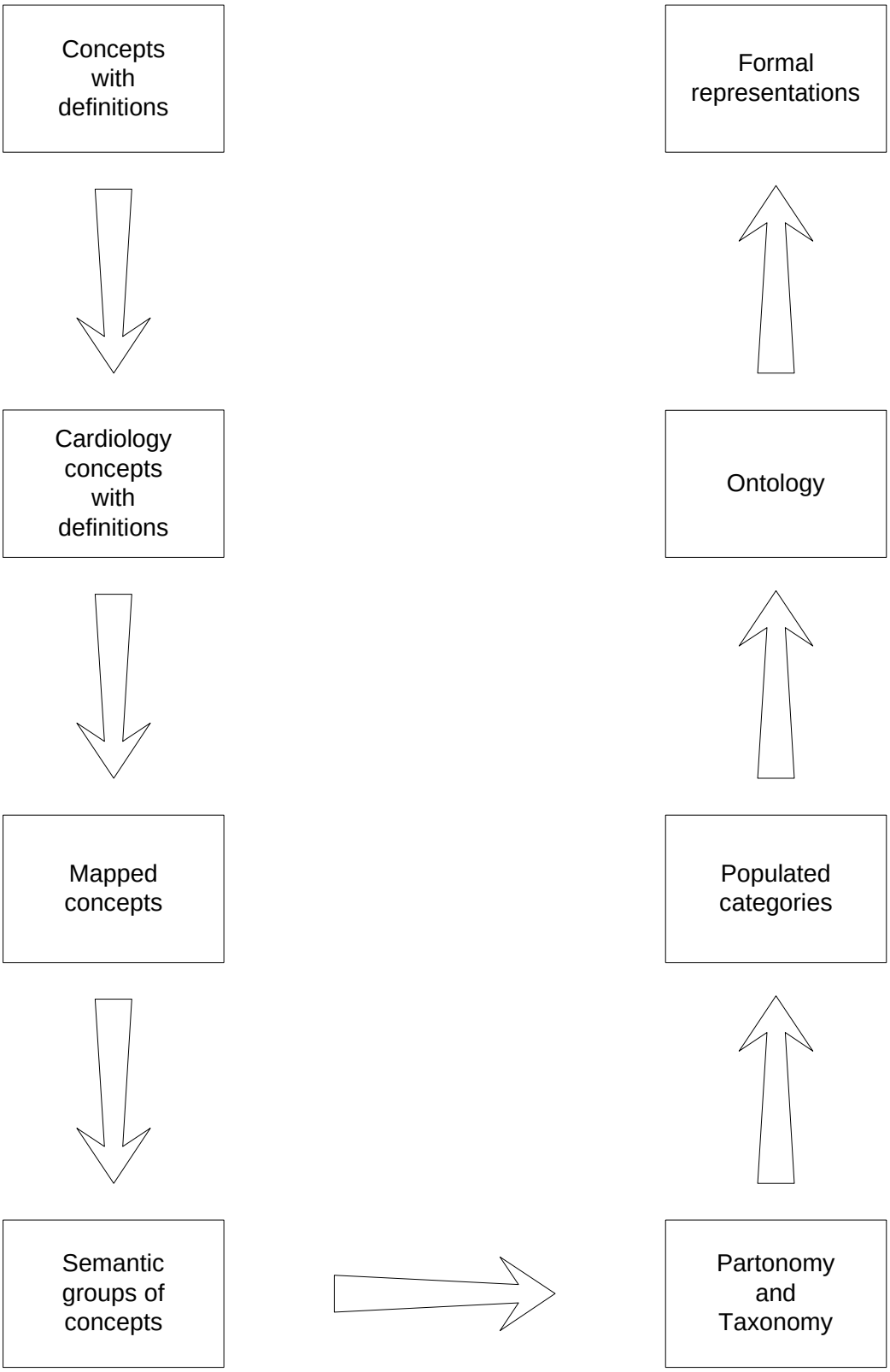
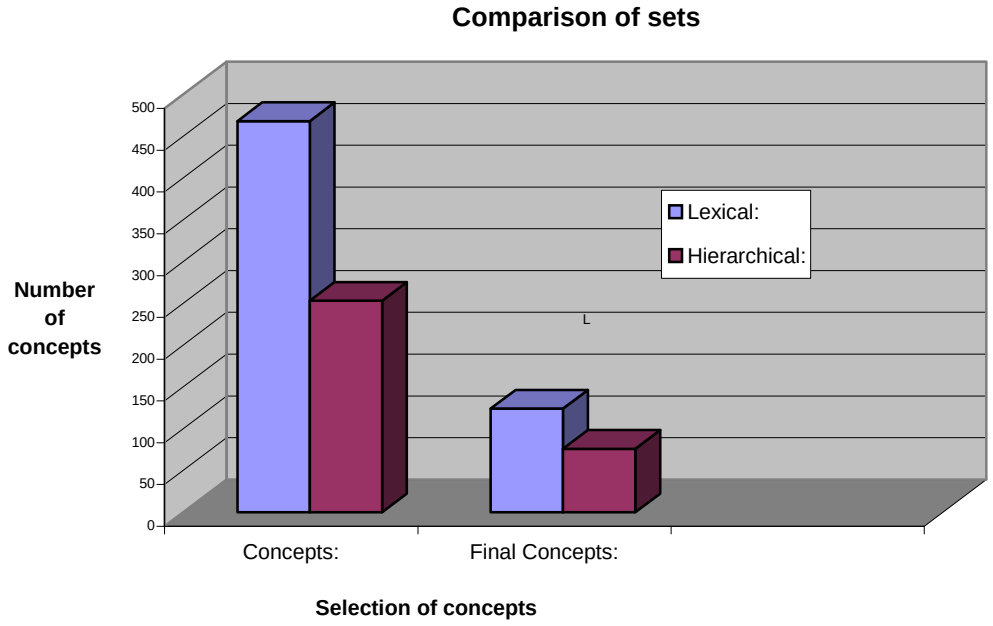


Table 1:

Comparison of sets:
(lexical-based and hierarchical-based)

	Concepts:	Concepts/MRDEF:	Final Concepts:	Final Concepts / MRDEF (Final Concepts / UMLS):
Lexical:	**468**	1.37 %	**124**	**0.36 %** (0.000155 %)
Hierarchical:	**253**	0.74 %	**76**	

Graph 2:



Graph 3:

Comparison of sub-sets of lexical and hierarchical methods

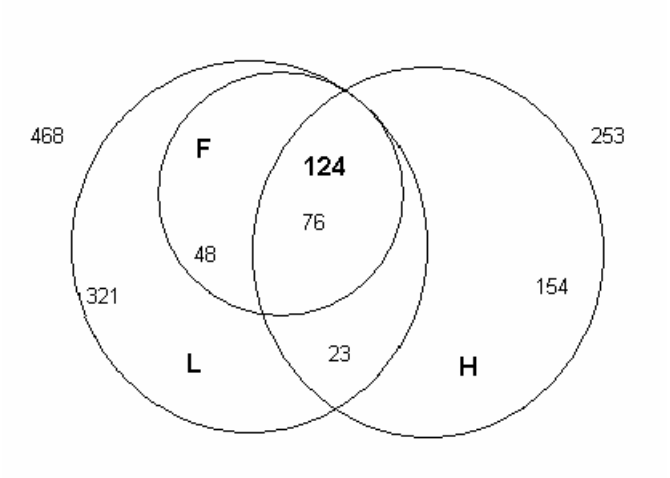


Table 2:

METAMAP EXTRACTED CONCEPTS:

Total:	POS:	NEG:
2152	1835	317
	85.27 %	**14.73** %

Table 3:

UNIQUE EXTRACTED CONCEPTS:

Total:	POS:	NEG:
987	804	183
	81.46 %	**18.54** %

Table 4:

Number of Concepts by Semantic Type:

Functional Concept	98	Injury or Poisoning	12	Educational Activity	3
Therapeutic or Preventive Procedure	91	Sign or Symptom	9	Body Space or Junction	2
Qualitative Concept	62	Research Activity	8	Cell	2
Spatial Concept	47	Substance	8	Cell Function	2
Body Part, Organ, or Organ Component	46	Body Substance	7	Human	2
Diagnostic Procedure	44	Pharmacologic Substance	7	Population Group	2
Finding	44	Mental Process	7	Neoplastic Process	2
Medical Device	37	Laboratory Procedure	6	Classification	1
Temporal Concept	34	Congenital Abnormality	5	Group	1
Quantitative Concept	31	Acquired Abnormality	5	Environmental Effect of Humans	1
Pathologic Function	30	Occupational Activity	5	Group Attribute	1
Organ or Tissue Function	28	Phenomenon or Process	5	Conceptual Entity	1
Manufactured Object	24	Biomedical or Dental Material	5	Idea or Concept	1
Natural Phenomenon or Process	23	Inorganic Chemical	4	Biologically Active Substance	1
Intellectual Product	21	Chemical Viewed Structurally	4	Lipid	1
Disease or Syndrome	21	Idea or Concept	4	Clinical Attribute	1
Spatial Concept	21	Anatomical Structure	4	Patient or Disabled Group	1
Body Location or Region	18	Body System	4	Group	1
Health Care Activity	16	Indicator, Reagent, or Diagnostic Aid	4	Biologic Function	1
Laboratory or Test Result	15	Daily or Recreational Activity	4	Chemical	1
Tissue	14	Organic Chemical	4	Embryonic Structure	1
Organism Function	13	Anatomical Abnormality	3	Individual Behavior	1
Organism Attribute	12	Animal	3	Age Group	1
Physiologic Function	12	Element, Ion, or Isotope	3		

Graph 4:

Agents and Processes of a Medical Act

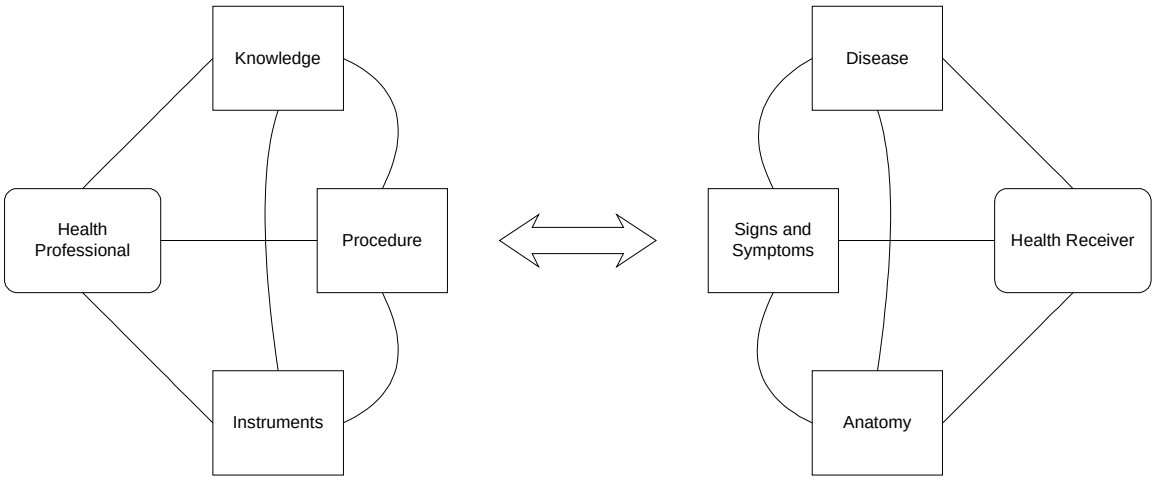


Table 5:

Comparison between former information model and the research categorization:

Original information model:	Research categorization:
Action	Action
Conditional Action	Conditional Action
Anatomical Site	Anatomical Entity
Pathology	Pathology
Instrument	Instrument
Measure	Measure
Sign	Phenomenon
Mode	Method
Circumstance	Time or Circumstance
Goal	Purpose
Expected Result	Result or Product
Parameter	-
-	Material or Energy
-	Physiology
-	Space
-	Receiver

Table 6:

Comparison between Semantic Groups and research categorization:

Semantic Groups:	Research categorization:
Procedures / Activities & Behaviors	Action
Procedures	Conditional Action
Anatomy	Anatomical Entity
Disorders	Pathology
Devices / Objects	Instrument
Concepts & Ideas	Measure
Phenomena	Phenomenon
Concepts & Ideas / Occupations	Method
Concepts & Ideas	Time or Circumstance
Concepts & Ideas	Purpose
Phenomena	Result or Product
Chemicals & Drugs / Objects	Material or Energy
Physiology / Phenomena	Physiology
Concepts & Ideas	Space
Living Beings	Receiver

Table 7:

Comparison between MAOUSSC and research categorization:

MAOUSSC:	Research categorization:
Nature	Action
Nature	Conditional Action
Topography / Matter / Access Pathway	Anatomical Entity
Diseases	Pathology
Instrument	Instrument
-	Measure
-	Phenomenon
Concepts & Ideas / Occupations	Method
-	Time or Circumstance
-	Purpose
-	Result or Product
Matter	Material or Energy
Biologic Process	Physiology
-	Space
-	Receiver

Table 8:

Number of concepts for each category:

Category:	Concepts:	Category:	Concepts:
Action	168	Conditional Action	5
Anatomical Entity	122	Method	19
Pathology	118	Phenomenon	30
Instrument	69	Measure	48
Material or Energy	43	Result or Product	12
Physiology	112	Time or Circumstance	31
Purpose	23	Space	50
-	-	Receiver	9

Table 9:

Comparison between human and automatic populating methods:

Author / Automatic	POS	NEG	
POS	797	172	969
	80. 75 %	**17.43** %	**98.18** %
NEG	7	11	18
	0.71 %	**1.11** %	**1.82** %
	804	183	987
	81.46 %	**18.54** %	

Table 10:

Number of Semantic Types for Types of Result of Populating Methods:

Type of Result:	H(+)-A(+)	H(+)-A(-)	H(-)-A(+)	H(-)-A(-)
# Semantic Types:	64	44	4	4

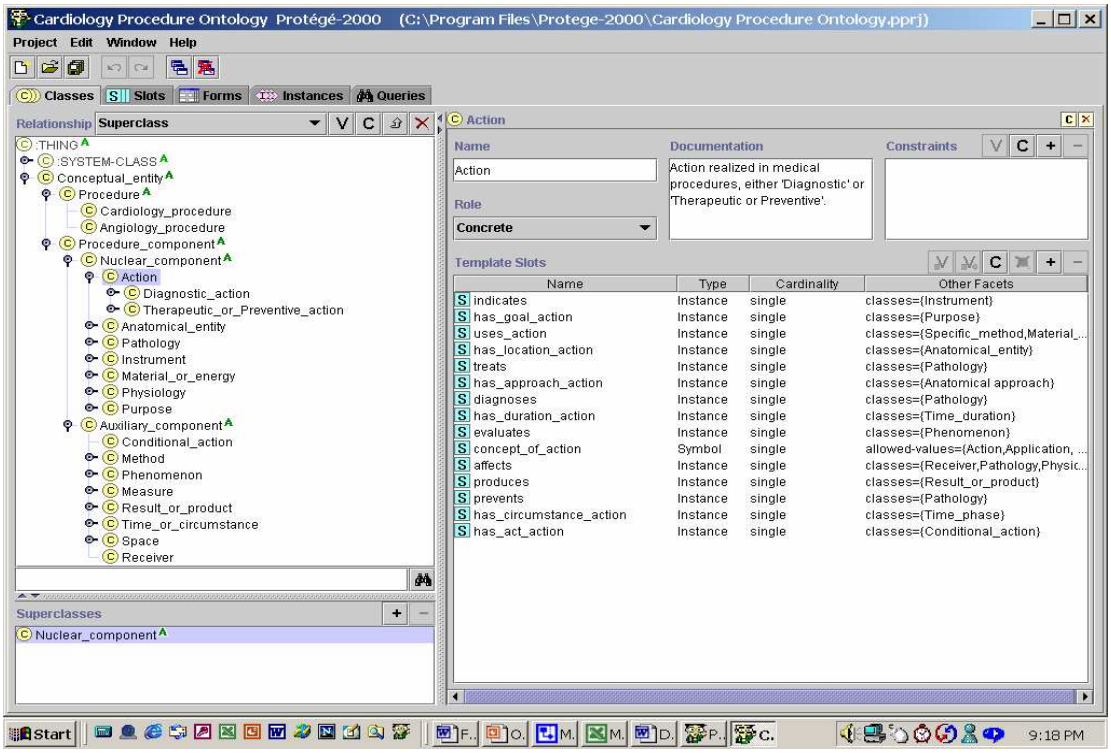
Table 11:

Number of equivalent and different Semantic Types for both methods:

>< =	H(+)-A(+)	H(+)-A(-)	H(-)-A(+)	H(-)-A(-)
H(+)-A(+)		0	43	0
H(+)-A(-)	4		0	1
H(-)-A(+)	1	4		0
H(-)-A(-)	4	3	4	

Figure 1:

Ontology in Protégé:

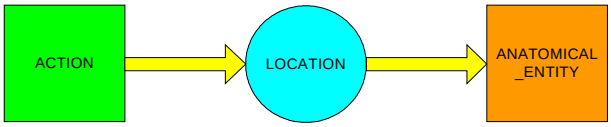


Graph 5:

Conceptual Graph 1:

Action has_location Anatomical_entity

[ACTION] > (LOC) > [ANATOMICAL-ENTITY]

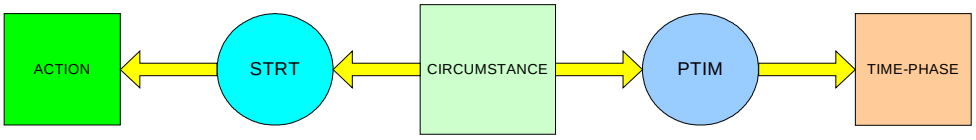


Graph 6:

Conceptual Graph 2:

Action has_circumstance Time_phase

[ACTION] < (STRT) < [CIRCUMSTANCE] > (PTIM) > [EVENT]

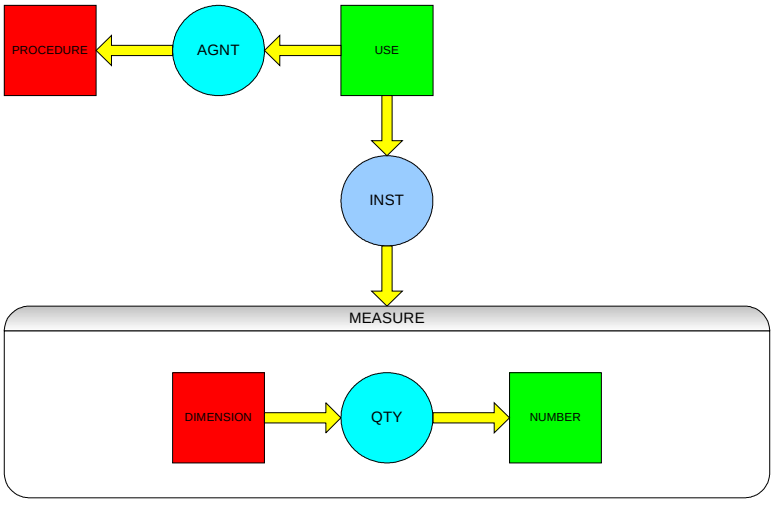


Graph 7:

Conceptual Graph 3:

Procedure uses Measure

[PROCEDURE] < (AGNT) < [USE] > (INST) > [MEASURE: [DIMENSION] > (QTY) > [NUMBER]]



Acknowledgements:

The following researchers have collaborated with the work developed in this research.
Susanne Humphrey, Dimitar Hristovski, and Anantha Bangalore.
My special thanks to Alexa T. McCray, Olivier Bodenreider, Tom Rindflesch, and Alan Aronson.

References:

1) Patel, V. Medical Cognition In: *Handbook of Applied Cognition*. Ed. Francis T. Durso. Chichester: John Wiley and Sons Ltd, 1999

2) Mayer, R. E. Thinking, Problem Solving, Cognition - 2nd edition. New York: W. H. Freeman and Company, 1992

3) Sowa, John F. Knowledge Representation – Logical, Philosophical, and Computational Foundations. Pacific Grove: Brooks/Cole, 2000

4) Sowa, John F. Conceptual Structures – Information Processing in Mind and Machine. Reading: Addison-Wesley Publishing Company, 1984

5) UMLS Knowledge Sources. 12th Edition. Bethesda: National Library of Medicine, 2001

6) Achour, S.L., Dojat, M., Rieux C., Bierling, P., Lepage, E. A UMLS-based Knowledge Acquisition Tool for Rule-based Clinical Decision Support System Development. *Journal of the American Medical Informatics Association*, 8(4): 351-360, 2001

7) Uschold, M. and Gruninger, M. Ontologies: Principles, Methods and Applications. *The Knowledge Engineering Review*, 11(2): 93-136, 1996

8) Simons, P. Parts – A Study in Ontology. Oxford University Press, 1987

9) CRISTAL'S MAOUSSC at http://noemed.univ-rennes1.fr/MAOUSSC/

10) Rosch, E. Principles of Categorization. In: *Concepts – Core Readings*. Ed. Margolis, E. & Laurence S. Cambridge: The MIT Press, 1999

11) Merriam-Webster's Collegiate Dictionary On-line. http://www.webster.com/